# Link Characteristics that Promote Clicks

STATCOM

6/18/2021

## Executive Summary

### Questions of Interest

- Which link characteristics (bolded, font size, font color, location within newsletter) lead to more clicks?
- Which topics mentioned in the links (e.g. animals, art, family, hunger) tend to attract clicks?

### Statistical Analysis

- Bar plots to visually display click response based on link characteristics
- Statistical model to examine how factors interact and support the trends shown in the bar plots
- Qualitative analysis on the words associated with a link

### Takeaways

- The location of the link in the newsletter affects how often it is clicked. Specifically, the further down the link is in the newsletter, the less likely it is to be clicked on. Therefore, we suggest placing the most important or time-sensitive links at the top of the newsletter.
- Links that are bolded get slightly more clicks than links that do not have bolded text.
- Links that are Mandarin Orange, Danube, or Denim colored get more clicks than other colors.
- Contrary to our previous report on the trends in click probability, having a picture associated with a link does not statistically significantly increase the probability that a link will be clicked. However, links with an image still receive slightly more clicks than links that do not have an image.
- Opportunities with baby chicks and sharing one's skills are significantly popular.

# Full Report

In this report, we investigate the characteristics of a links that make it more likely to be clicked on. We focus on newsletters from January 2019 to December 2020.

The rest of the report is organized as follows. First, we give a brief description of how the data was obtained and a synopsis of the assumptions we made to analyze the click data. Then, we introduce the features used in the model and analyze how click rates were affected by these features separately. Finally, we fit a statistical model to the data and interpret the results.

## Data Description

The data comes from a few sources: the CSV and raw text files generated from iContact and the HTML source code from the links in each of the newsletters. From the CSV files, we determine the unique number of times a link was clicked on. We define a unique click to be a unique combination of subscriber ID, newsletter date, and link; in other words, if a subscriber clicked on the same link from the same newsletter, we do not count that click. We also identify the time of day the newsletter was sent and whether it was before or after the COVID-19 pandemic was declared (03/20/20) from the CSV files. The raw text files are used to get an approximation of how far down the newsletter the link is, e.g. a link that is about half-way down the newsletter would be estimated around 50%, and the text associated with a link. Finally, we obtain style characteristics and whether the link was an image or had an image associated with it from the HTML source code.

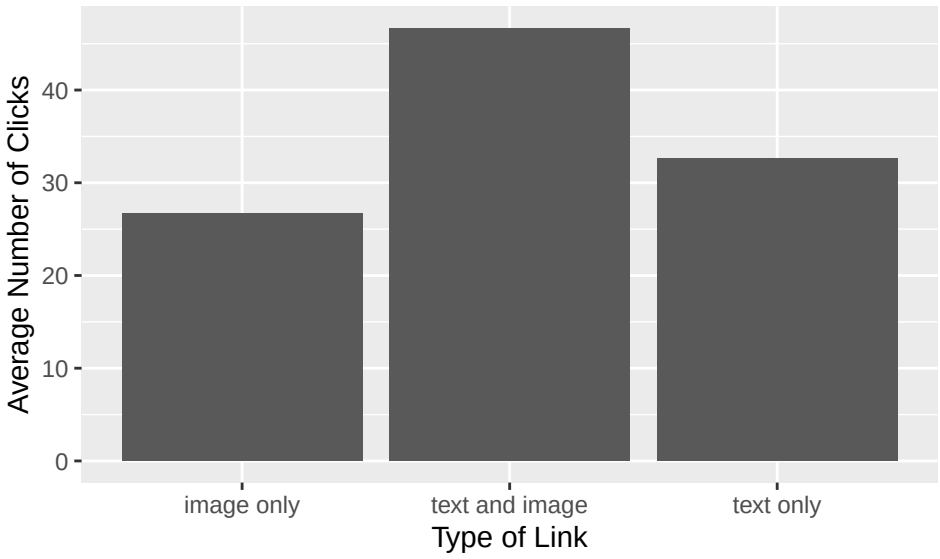
Additionally, we only know the click count for each link address, not the actual link, within a newsletter. Thus, if there are multiple links with the same address in a newsletter, we don't know how many clicks each of those links received. To alleviate this issue of duplicate links, we assume that the first text link with a given address received all the clicks associated with that address. We distinguish between unique links (those with addresses that appear only once) and duplicated links (those with addresses appearing multiple times) in our model.

Below is a summary of the features we created for text links:

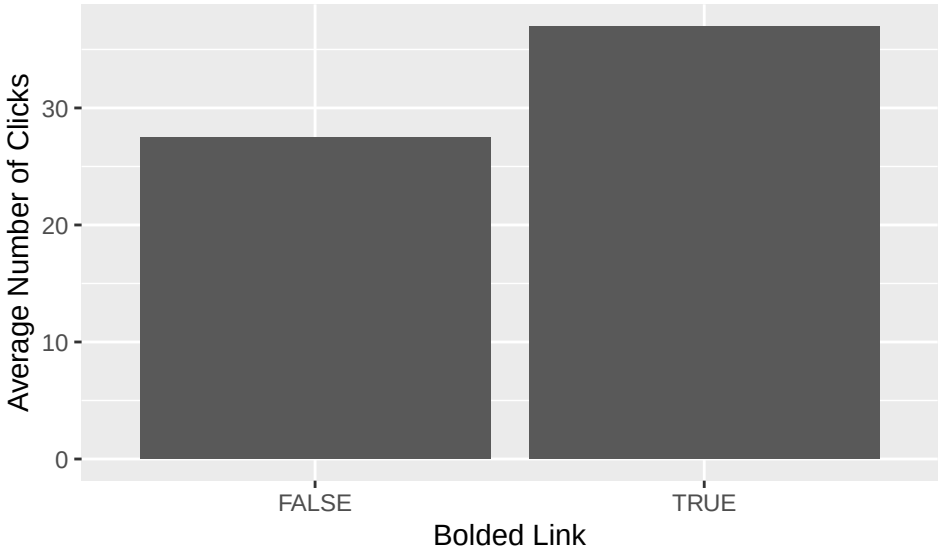
1. Bolded: whether the image is bolded.
2. Font Size: ranges from 10-48 point.
3. Font Color: 26 possible colors.
4. Image Associated: indicator for whether there is an image within the newsletter with the same link address.
5. Hour: hour of when the newsletter was sent
6. COVID: indicator for whether the COVID-19 pandemic was underway
7. Location within document: cumulative percentage down the document a link is
8. Duplicate: indicator for whether the link was used more than once in a newsletter

## Data Exploration

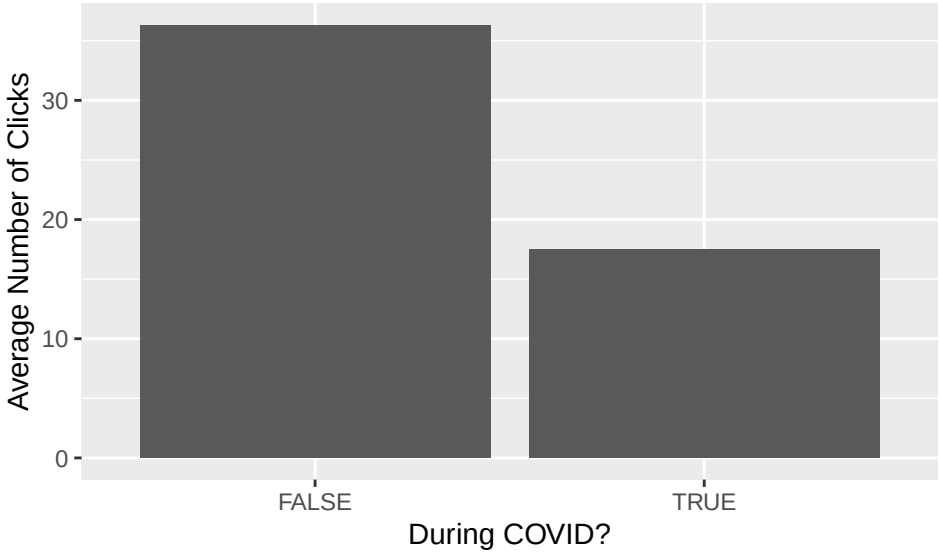
Before fitting any models to the data, we explore how the number of clicks a link received depended on the variables mentioned above. We define the average number of times a link was clicked as the number of times the link was uniquely clicked, defined above, divided by the number of newsletters the link appeared in. It is important to note that in doing this, we do not control for how many times a link was used within the same newsletter. For each of the categorical variables, we graph the category and the average number of times a link was clicked below.




In the bar plot above, the label 'image only' refers to links that only had a picture associated with it, 'text only' refers to links with only associated text, and 'text and image' refers to links with both an image and associated text. Based on the bar plot, it appears that a combination of text and pictures encourages people to click on a link.

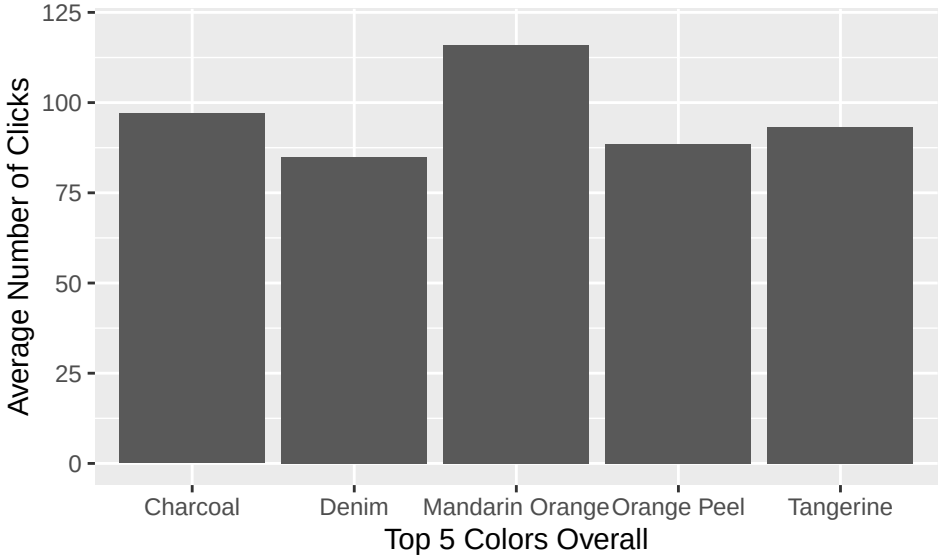



Based on the bar plot above, it appears the bolding the text associated with the link also increases the chance that someone clicks on it.

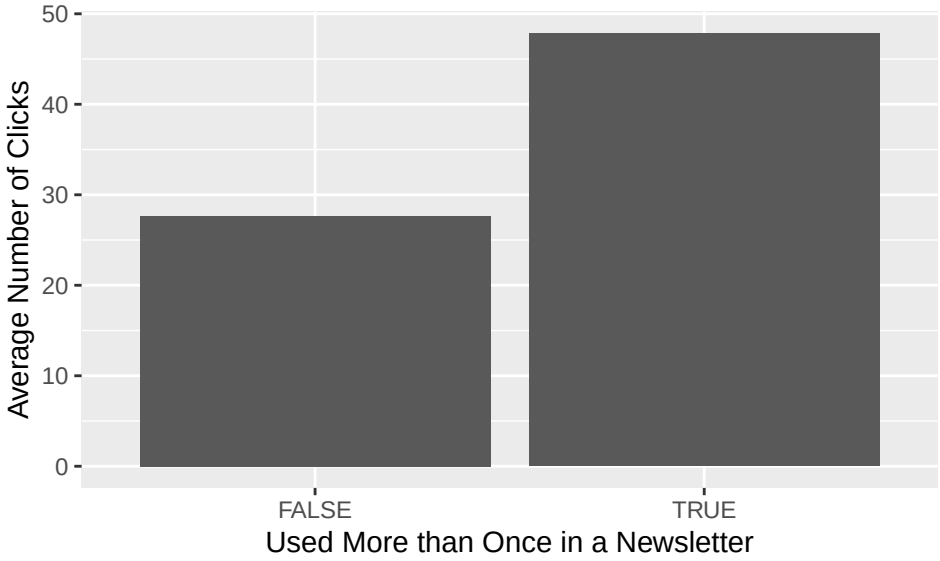



Based on the bar plot above, it seems COVID impacted how much subscribers choose to interact with the newsletters. This is not all that surprising given the challenges everyone was facing during the pandemic.

In the next two bar plots, we focus on the top five text color choices across all newsletters and the top five text color choices that appeared in more than one newsletter.




Any color of orange seems to grab people's attention! Mandarin Orange only appeared in the newsletter promoting the Remote Volunteer Project: DIY Family Essentials Kits opportunity so it is tempting to think the large number of clicks this color received may have more to do with the highly-relatable project. However, this project was advertised in four different newsletters using links colored as cinnabar and falu red (both are different tints of red) and these other newsletters had less than 85 clicks each. While there are more factors at play than just the link color, the fact that the newsletters advertising the same opportunity in red got fewer clicks suggests that a text color of orange is more impactful.




Finally, the bar plot above indicates that links that were used multiple times in a newsletter were clicked more often. This is somewhat surprising since interested individuals may click the first link they come to, but the plot above suggests more exposure to the opportunity encourages participating. However, the model below suggests this may not be the case.

## Model Fitting

We attempted to fit a linear regression, a Poisson regression, and a Negative Binomial regression to the click data with no success. However, we found that if we divided the number of clicks by the number of subscribers each newsletter was sent to, a zero-inflated beta regression worked quite well. The beta regression allows us to model proportion data (data that's bounded between zero and one, non-inclusive); the "zero-inflated" in the name refers to extending the beta regression to include observations with a value of zero. The zero-inflated beta regression fits three parameters: mu, sigma, and nu. The mu variable corresponds to the mean of the click proportion (relative to the number of subscribers) and is modeled in a similar manner to simple linear regression.

The variables in our model are the following: doc_prop, bolded, color, font_size, hour, covid_ind, imag_assoc, and dup. "doc_prop" is the proportion down the document a link is; in other words, a link that is about halfway down a newsletter will be about 50%. "bolded" indicates whether a link was bolded. "color" is the color of the link as determined by https://www.color-blindness.com/color-name-hue/.

Below we give a histogram of click proportion and the fitted model parameters for mu. From Table 1 below, we see that where the link is in the document, whether the link is bolded, and the color of the link make a statistically significant difference on whether the link is clicked or not. Additionally, we see the top five colors shown above are also statistically significant as well as Charcoal and Falu Red. Interestingly, the indicator for whether the variable is duplicated or not is not significant in the model. This lack of significance may be because the location of the link in the document also somewhat captures this information.

Finally, the mu coefficients given in the table below are, unfortunately, uninterpretable in their raw form. Luckily, a transformation of these coefficients gives the odds ratio of each variable. For the variables that are statistically significant at or below the 0.001 level, we give the odds ratios in Table 2. The odds ratio is interpreted as follows: When the location of the link in the document increases by one percentage point the odds of it being clicked is 0.46 times the odds of it being clicked in the original position. Overall, an odds ratio greater than one indicates a positive association and an odds ratio less than one indicates a negative association; note, these agree with the signs of the coefficients in Table 1.

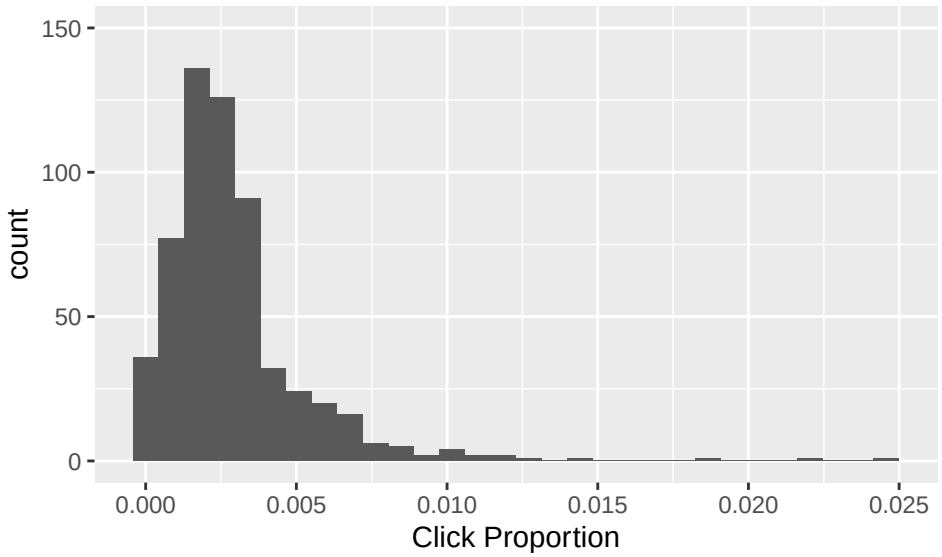

150
100
50
0
count
0.000
0.005
0.010
0.015
0.020
0.025
Click Proportion

**Table 1: Mu Coefficients**

| | Estimate | Std. Error | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|
| (Intercept) | -5.96 | 0.26 | -23.18 | < 2e-16 | *** |
| doc_prop | -0.78 | 0.14 | -5.56 | 0.00 | *** |
| boldedTRUE | 0.22 | 0.06 | 3.50 | 0.00 | *** |
| color_nameBlack | 0.61 | 0.36 | 1.71 | 0.09 | . |
| color_nameBlack Pearl | -0.09 | 0.44 | -0.21 | 0.84 | |
| color_nameCharcoal | 1.06 | 0.31 | 3.39 | 0.00 | *** |
| color_nameChocolate | 0.47 | 0.24 | 1.93 | 0.05 | . |
| color_nameCinnabar | 0.07 | 0.33 | 0.23 | 0.82 | |
| color_nameCitron | 0.13 | 0.30 | 0.45 | 0.66 | |
| color_nameDanube | 1.62 | 0.34 | 4.72 | 0.00 | *** |
| color_nameDenim | 1.49 | 0.31 | 4.84 | 0.00 | *** |
| color_nameDim Gray | -0.23 | 0.73 | -0.32 | 0.75 | |
| color_nameEastern Blue | 0.28 | 0.25 | 1.09 | 0.28 | |
| color_nameEclipse | 0.37 | 0.24 | 1.58 | 0.12 | |
| color_nameFalu Red | 0.64 | 0.24 | 2.70 | 0.01 | ** |
| color_nameGamboge | 0.15 | 0.53 | 0.29 | 0.77 | |
| color_nameGrey | 0.18 | 0.24 | 0.77 | 0.44 | |
| color_nameMandarin Orange | 1.35 | 0.38 | 3.52 | 0.00 | *** |
| color_nameMariner | 0.26 | 0.59 | 0.43 | 0.67 | |
| color_nameNero | 0.18 | 0.68 | 0.27 | 0.79 | |
| color_nameOrange Peel | 1.04 | 0.34 | 3.08 | 0.00 | ** |
| color_nameSlate Blue | 0.45 | 0.61 | 0.75 | 0.46 | |
| color_nameTangerine | 0.94 | 0.28 | 3.42 | 0.00 | *** |
| color_nameTeal | 0.44 | 0.24 | 1.82 | 0.07 | . |
| color_nameTenne | 0.85 | 0.44 | 1.92 | 0.06 | . |
| color_nameTyrian Purple | 0.28 | 0.55 | 0.51 | 0.61 | |
| color_nameWhite | 0.42 | 0.24 | 1.72 | 0.09 | . |
| font_size | -0.01 | 0.01 | -1.03 | 0.31 | |
| hour | 0.01 | 0.01 | 1.29 | 0.20 | |
| covid_indTRUE | -0.70 | 0.08 | -9.17 | < 2e-16 | *** |
| image_assocTRUE | 0.03 | 0.10 | 0.32 | 0.75 | |
| dupTRUE | 0.12 | 0.09 | 1.30 | 0.19 | |

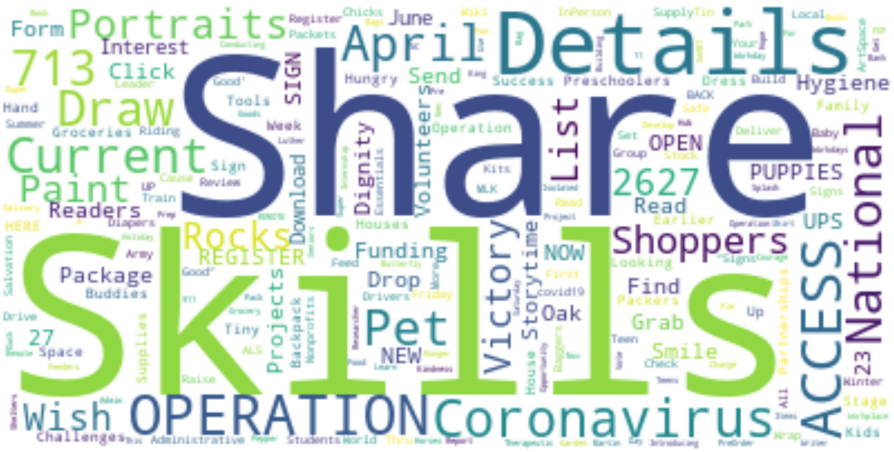
**Table 2: Odds Ratio**

| | Estimate | Odds Ratio |
|---|---|---|
| (Intercept) | -5.96 | 0.003 |
| doc_prop | -0.78 | 0.460 |
| boldedTRUE | 0.22 | 1.240 |
| color_nameCharcoal | 1.06 | 2.884 |
| color_nameDanube | 1.62 | 5.061 |
| color_nameDenim | 1.49 | 4.445 |
| color_nameFalu Red | 0.64 | 1.896 |
| color_nameMandarin Orange | 1.35 | 3.845 |
| color_nameOrange Peel | 1.04 | 2.817 |
| color_nameTangerine | 0.94 | 2.571 |
| covid_indTRUE | -0.70 | 0.495 |

### Qualitative Text Analysis

Finally, we explored what words encouraged subscribers to click on a link by creating word clouds. The word clouds are composed of the capitalized words that were contained in each link. (The requirement for capitalization was to eliminate filler words.) the size of the words corresponds to the proportion of unique clicks relative to the number of total clicks a newsletter obtained. Note, this is slightly different than the proportion of clicks defined above for the zero-inflated beta model.

The first word cloud below is for any links that were not social media for Activate Good.




The second word cloud below is for links that were for opportunities as defined by the link containing "opportunity" in the address; these links correspond to volunteer opportunities for subscribers.


As we can see, opportunities with baby chicks and sharing one's skills are significantly popular compared to other opportunities!